A Model of Clinical Query Management that Supports Integration of Biomedical Information Over the World Wide Web

William M. Detmer and Edward H. Shortliffe

Section on Medical Informatics, Stanford University School of Medicine

A model of clinical query management is described that supports the integration of various types of biomedical information and the delivery of that information through a common interface. The model extends the architecture of the World Wide Web to include a Common Gateway Interface (CGI) mediator, which takes in user queries, performs syntactic and semantic processing to transform the input to a canonical form, selects the appropriate information sources to answer the query, translates the canonical query statement into a query of each information resource, queries the chosen information sources in parallel, and controls the analysis and display of results. We describe WebMedline, a CGI mediator that implements portions of this model, and discuss the benefits and limitations of this approach.

INTRODUCTION

Studies have shown that information needs that arise in clinical practice are frequently unmet [1, 2]. Some of the barriers to satisfying these needs include lack of up-to-date information resources, poor organization of available information, ignorance of the availability of relevant information, lack of time for searching, and uncertainty about the scientific validity of published information [1].

Progress could be made if up-to-date information relevant to the need were rapidly available at the point of care. Electronic resources such as bibliographic databases, on-line textbooks, and drug databases provide an attractive means of delivering such information because of the speed by which these sources can be searched. In the era of the Internet, additional advantages are realized when information sources are available over wide-area networks: centralized maintenance of content, rapid updates of information, and distribution of cost across larger user communities. Using the World Wide Web architecture, further benefits accrue because information can be displayed in hypermedia through a common interface on a variety of computing platforms.

However, there still exist many barriers to the use of network-based electronic information for answering clinical questions. The first barrier is computer literacy. The second barrier is the lack of knowledge of available resources and which one is best suited to answer a particular question. The third barrier is the difficulty in learning the user interface for each information resource, including where the resource is located, how to issue queries in the specialized search language, and how to manage the display of results. Consequently, most practicing clinicians do not have the knowledge and experience needed to translate an information need quickly into a sophisticated query of an appropriate database. Therefore, models of clinical query management are needed that support the expression of information need and that orchestrate the retrieval, integration, and display of information from disparate network-based information sources.

BACKGROUND

The World Wide Web

The World Wide Web architecture was developed by Berners-Lee [3] and is based on a generic object-oriented protocol, the Hypertext Transfer Protocol (HTTP). This protocol manages requests in the form of a Uniform Resource Locator (URL) and delivery of information as a Multipurpose Internet Mail Extension (MIME) objects. The most common objects delivered by the HTTP protocol are documents written in the Hypertext Markup Language (HTML), a subset of the more general Standard Generalized Markup Language (SGML) [4]. HTML adds structure to ASCII text documents, and WWW browsers (such as Mosaic or Netscape) use this structure to display the text in a graphical manner. Beyond designating the structure of documents, HTML provides a syntax for embedding graphics, images, sounds, and video, as well as hyperlinks to other documents [3].

One of the main tenants of the HTTP protocol is that it is stateless: after the HTTP server returns the requested information, the session is terminated. No information about the state of the user is maintained. Because many interactive processes require maintenance of state information, developers have maintained state information in hidden fields of HTML forms or in databases resident on the server.

To support the processing of user input, the Common Gateway Interface (CGI) standard was developed. This standard assures that WWW browsers, HTTP servers, and external processes communicate using a standard set of parameters. When a hyperlink or HTML form is used to initiate a CGI process, the HTTP server receives the request, starts the CGI process with the parameters submitted by the user, waits for the output of the CGI, and delivers the output to the browser. The CGI application can use the supplied parameters to perform almost any task: make a database query, annotate a document, or send an electronic mail message.

Information integration

The goal of integrating biomedical information for clinicians is not a new one. In fact, the National Library of Medicine's (NLM) Unified Medical

0195-4210/95/$5.00

Language System (UMLS) was developed with this as one of its main goals [5]. Several groups have used the UMLS as the centerpiece of their information-integration models [6, 7]. Some other projects, such as [8], have used the UMLS to created links between clinical data and relevant bibliographic information. Investigators in the fields of database theory [9] and artificial intelligence [10] have also proposed models that support data and information integration. What distinguishes our model from other work is its emphasis on integration of network-based biomedical information and its support for display of information in hypertext.

METHODS

The general architecture of the clinical query model is composed of five components (Figure 1): (1) a WWW Browser, (2) an HTTP server, (3) a CGI mediator, (4) a representation of medical concepts, and (5) information resources accessible over wide-area networks. The steps required to issue a query and to display the results include: the user fills in an HTML form, submits the contents of the form, and waits for the output of the CGI process to be displayed by the browser. The CGI mediator performs six tasks, as shown by the bullets in Figure 1. Each task of the CGI mediator is discussed below, after the discussion about user input.

User input

The user interface is constructed using HTML forms. Although the general model does not require a specific type of interface, it is desirable that the interface constrain input enough to facilitate its semantic processing. One way to constrain user input while maintaining expressivity is to provide templates that encode generic types of queries. As has been pointed out by other authors [6,11,12], most clinical questions map to a limited set of query types. This knowledge can be used to develop input templates that allow expression of clinical questions.

Other characteristics that can aid in semantic processing are characteristics of the user, such as level of medical training, specialty, and degree of experience with computers. This user profile helps to select the best information source to answer the user's question and infer the best way to summarize and display results. This information can be obtained during an initial user session, and then be retrieved from a user-profile database when additional sessions are initiated.

Another useful characteristic is the goal of the user's question. Prior authors [12,13] have enumerated common types of goals, but have done so in a limited domain. More extensive investigation is required to identify a more general set of goal types. Such information can help narrow the scope of the query.

Syntactic processing

Once the CGI mediator receives the user input, it performs syntactic processing. Extraneous characters and stop words are removed and stemming algorithms are used to derive word roots from input.

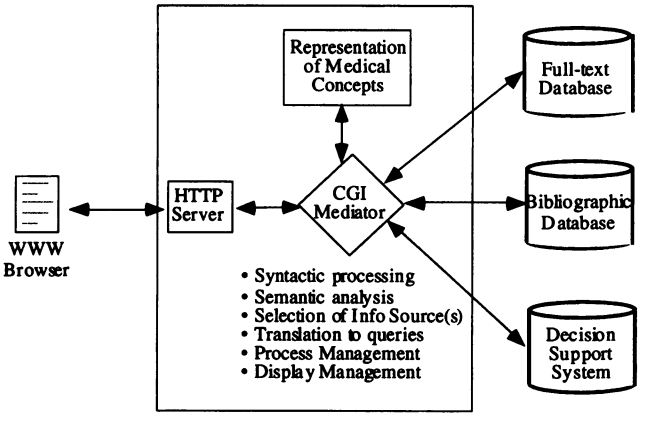

Figure 1. The model for clinical query management

Semantic processing

In this step the input is processed to identify semantic content. Synonyms are mapped to a single concept identifier and object types for each concept are identified or inferred. The user profile and purpose of the query are used to support this semantic analysis.

Required for semantic analysis is a representation that contains medical concepts, their semantic types, and their inter-relationships. The representation can be used to identify more general and specific concepts, which provides a mechanism for creating a more or less general canonical query representation. A representation that would support this model is the UMLS [5]. The UMLS Metathesaurus can be used to map query strings to concepts and the Semantic Network can aid in inferring relationships among concepts. An anticipated problem with using the UMLS is the current lack of depth of concepts for clinical findings [14]. Other representations that contain more knowledge of concepts and relationships but have narrower scope could also provide support for this model.

Selection and prioritization of information sources

The user-profile and purpose of the query, together with the canonical query representation, are used to infer how the information sources should be selected and prioritized. Examples include rules such as those in the UMLS Information Sources Map [5], probabilistic networks [13], or even neural networks.

Translation of concepts into queries

Once concepts and relations have been identified from the user input and the information sources have been selected, the concept must be translated into the query language of each of the chosen information sources. Separate translators are created for each information resource to allow resources to be added and removed without disruption to the entire system. Translators can be implemented by using the UMLS to map the query in its canonical form to the source vocabulary, if the source vocabulary is represented in

the Metathesaurus. If the information source is not represented in the Metathesaurus but can be queried by full-text algorithms, the Metathesaurus can be used to construct a query statement using the synonyms and lexical variants of its core concepts [5].

Process management

Once the information sources have been chosen and the query statement has been composed, the CGI mediator queries each of the individual information sources. Required tasks include opening a session with each information source using an agreed upon protocol, navigating the logon sequence, issuing the query statement, designating the display format, collecting the output, and closing the session. When more than one resource is being queried at a time, simultaneous processes are initiated and these processes are coordinated so that their output is not intermingled.

Display management

Once the results of all processes are available, the display manager performs analysis of the retrieved information. Analysis is necessary if the display of output is to be different than the output from the individual information resources. For example, the output from an information system that displays information in reverse chronological order, might be analyzed for the number and location of the query terms in the output. This analysis could be used to resort or filter the output

Once analysis of output is completed, the display manager controls how information from each resource is summarized, how information is integrated across resources, and how character-based information is transformed into hypertext

RESULTS

To explore the general model summarized above, we have developed an application named WebMedline, a hypertext interface that replaces a standard, character-based Medline interface. In an experimental version, WebMedline retrieves citations from the Medline database and integrates them with critical reviews of these citations published by the American College of Physicians in the *ACP Journal Club*. The WebMedline Mediator parses user input submitted via an HTML form, queries both Medline and the *ACP Journal Club* databases, integrates the output of the two searches, and displays the results in hypertext.

WebMedline

WebMedline is a hypertext interface to Melvyl Medline, a database maintained by the University of California [15]. The standard method of accessing Melvyl Medline is to initiate a terminal-based Telnet session, logon to the Melvyl host, navigate the opening prompts, issue queries in the Melvyl Medline query language, and display results using a specialized language.

In contrast, WebMedline moves Medline to the World Wide Web by providing an HTML form, which the user fills out and submits for processing (Figure 2). Users enter text into predefined fields, such as Author, Title, Journal, and Keyword. In addition, they can choose from pop-up menus the database years, the between-field Boolean operator, the display type, and number of citations to retrieve. Finally, they can choose to constrain the search by standard limiters such as "English only," "Human subjects only," and "Publication Type" (e.g., randomized controlled trial).

When the contents of the HTML form are submitted, they are passed by the HTTP server to the CGI mediator. The mediator first performs syntactic processing by removing stop words and by stemming words to their roots. The WebMedline mediator performs semantic analysis in a limited fashion by querying the thesaurus function built into Melvyl Medline. Keywords entered in the HTML form are mapped to possible MeSH by a synonym-lookup algorithm. No automated mechanism currently exists to resolve ambiguities between words that have more than one meaning. Instead, the related MeSH terms are displayed to the user and the user chooses the term that best represents his intention. The current result of the syntactic and semantic processing is a query

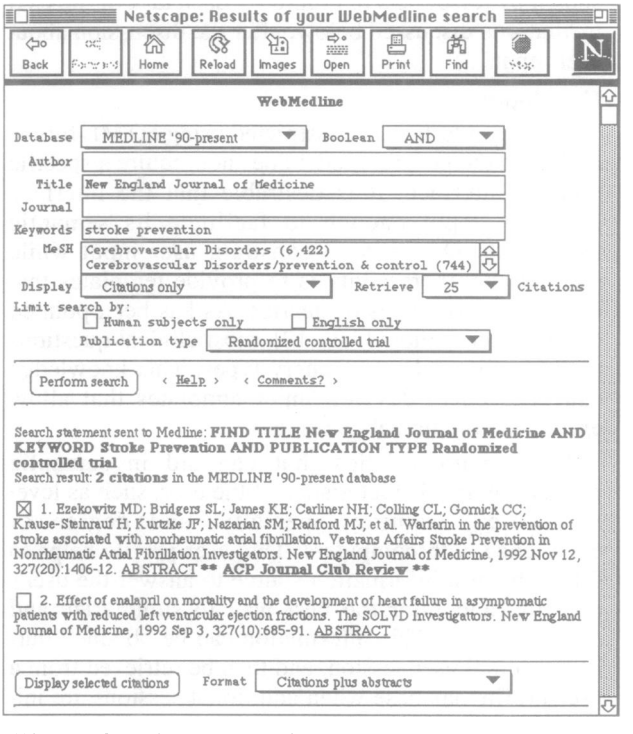

Figure 2. The results of a WebMedline search. The top portion of the page shows a new HTML form, which the user can use to further refine the query. For instance, the user can choose one of the MeSH terms that was returned as a by-product of the prior search. The bottom portion of the page contains the results of the search. Note that if a Medline or *ACP Journal Club* abstract exists for a particular citation, a hyperlink has been created dynamically. Also note that a user can select a number of citations by checking the box in front of the article and redisplay the selected citations in another form (e.g., with their abstracts).

representation that is more canonical than the original input, but is not the most general canonical representation envisioned by our model.

The translation process then begins: field descriptors are added to the values entered in particular fields and a Boolean statement is created by appropriate nesting of Boolean phrases. Once the information resource has been chosen (in this case, deterministically) and the query statement has been composed, the process manager proceeds to query each information resource. For the interaction with the Melvyl Medline database, WebMedline establishes a Telnet session with the Melvyl host, navigates the logon procedure and the opening prompts, issues the query statement, requests information in a particular format, and closes the Telnet session. It then parses from the Melvyl Medline output the NLM unique identifier for each citation and queries the *ACP Journal Club* database for reviews that have the same unique identifier. The query results from the two sources are then passed to the display manager.

The display manager performs analysis of the combined output, integrates the contents, and marks up the character-based output in hypertext. For the "citation only" display format, the mediator creates an "Abstract available" hyperlink when it encounters a flag that indicating a Medline abstract exists for the citation. In addition, if an *ACP Journal Club* review exists for a particular citation, a dynamic hyperlink is created. Because the HTTP protocol is stateless, the hyperlink must identify the source of the information and a unique identifier or path to the information to assure that the information can be retrieved during a new session with the information source.

WebMedline has undergone extensive testing at the University of California, San Francisco and Stanford University. During a 4-month testing phase, we made WebMedline available by word of mouth to users on the two campuses. During this trial period, we captured log data that included the user ID, computer address, date and time of the request, query statement, number of citations returned, number of MeSH term retrieved by keyword lookup, display type, and goal of the search. From this data a complete trace can be made of each session.

Preliminary analysis of the log data shows that users performed 12,482 searches during the trial period. The average time to perform a query was 10.1 seconds. A WebMedline "session," the number of consecutive searches of the database by a particular user, averaged 4.7 queries with a range of 1 to 26 queries, and lasted, on average, 3 minutes and 40 seconds. Of those who used WebMedline once, 48% returned to run a search at a later date. By the end of the 4-month trial, 10% of Medline sessions initiated from the Stanford campus used the WebMedline interface. With further analysis of the log data and correlation with satisfaction measures we expect to gain further insight into how users navigate through the WebMedline searching process, and how well the results meet the user's information need.

DISCUSSION

Integration of biomedical information using the World Wide Web architecture offers many advantages. First, WWW browsers, HTML forms, and CGI mediators allow simplified access from a variety of platforms to searchable information scattered across the Internet. Particularly important for busy clinicians, the WWW architecture provides a mechanism for shielding users from needing to know details of Internet protocols or the location of information resources. It also supports, through CGI applications like the WebMedline mediator, the ability to shield users from needing to know the query language of each information source. Instead, developers can design user interfaces that more closely match the needs and abilities of the user and let CGI mediators map data entered into the interface to the language of each information resource.

An additional advantage of the WWW architecture is the potential for integrating information that resides in disparate locations, that is structured in different ways, and that is searched by different information-retrieval methods. At a time when medical information is migrating from legacy systems to more structured databases, this architecture provides glue that can bring together information that resides in different types of environments.

An additional advantage is the ability to display the results in hypertext. The power of hypertext is that it enables hierarchical browsing and filtering; in an initial display of information, a summary can be presented with hypertext links to more detailed or related information. WebMedline takes advantage of hypertext capabilities by presenting citations in their most basic form and allowing users to explore Medline abstracts or *ACP Journal Club* reviews by following hyperlinks.

WebMedline and the Web-based architecture that support it are not without their disadvantages, however. Limitations of the interface-building facilities, absence of a standard method for maintaining state information, and lack of unique identification of documents constrain what can be done using the current instantiation of the architecture.

Although the HTML syntax allows creation of forms for various types of user input, it has considerable limitations. First, only a small number of interface objects are currently supported, and no type-checking is possible within a particular field. Second, no facility exists for creating dependencies between interface objects. For instance, if a user chooses a particular database in WebMedline, there is no way to change other interface objects to reflect those options available only for that particular database. The result is that if a form is be used to submit queries to several databases, it must contain interface objects that represent the intersection of all objects that would be used for each individual

information source. If the databases contain very different types of information, then the intersection of interface features could be small enough to make the common interface useless.

Another limitation of the current architecture is lack of a standard method for maintaining state. For example, when a user submits a query to WebMedline, the user receives back as one document both citations marked up in hypertext as well a copy of the original HTML form. The HTML form shows the standard interface objects, but also inserts in the fields the information that was entered for the previous search. This feature allows users to refine their query without entering all the original data. To accomplish this using the current architecture, state information is inserted in each version of the HTML form. Some information is entered as default information in the fields, while other information is embedded using hidden fields. This method of maintaining state in inefficient, because all the state information needs to be carried back and forth over the network. A more efficient approach is to maintain some state information in a database residing on the server. Principled methods for saving such state information need to be developed.

Finally, the integration of information from disparate resources can only be accomplished if standardized, unique document identifiers are used by information providers. Without such standard identifiers, incompatible naming conventions will proliferate and limit the potential for inter-resource linking. As an example of the difficulty linking resources without standard identifiers, we had to add NLM unique identifiers to each *ACP Journal Club* document in order facilitate the creation of hyperlinks from Medline to the *Journal Clubs*. Text-processing routines were needed to parse from the *Journal Clubs* the author, title, and journal of the article being reviewed. These parameters were then used in a batch process to lookup and store the NLM unique identifier in each *Journal Club* document. Such brute-force methods of integration could be avoided if information scientists and medical publishers work together to establish standards that specify how unique identifiers should be assigned to and embedded in medical documents. Proposed standards such as the Serial Item and Contribution Identifier (SISAC) forwarded by he American National Standards Institute [16] may be a useful starting place for creating such a standard.

CONCLUSION

We have developed a model for clinical query management that supports the integration of network-based information resources. The model supports the asking of clinical questions of a single, simplified interface and the prioritized retrieval and display of information from disparate, widely distributed information sources. The lessons learned during both the building of WebMedline and the integration of WebMedline with the *ACP Journal Club* reviews, point to further research that is required to achieve the goal of delivering relevant information to practicing clinicians at the point of care. Further work in medical knowledge representation that support clinical queries, extension of the HTML syntax, creation of a standard to manage state information, and standards for unique identification of documents are all required to meet this goal.

Acknowledgments

This work was conducted with the support of the National Library of Medicine under grant LM-07033. Computing facilities were provided by the CAMIS Resource, LM-05305. Dr. Shortliffe is supported by the Henry J. Kaiser Family Foundation under grant #84R-2459-HPE. The *ACP Journal Club* reviews were made available as part of a research collaboration between the American College of Physicians and several academic institutions.

References

1. Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? Ann Intern Med 1985;103(4):596.

2. Osheroff JA, Forsythe DE, Buchanan BG, Bankowitz RA, Blumenfeld BH, Miller RA. Physicians' information needs: analysis of questions posed during clinical teaching. Ann Intern Med 1991;114(7):576.

3. Berners-Lee T, Cailliau R, Luotonen A, Frystyk Nielsen H, Secret A. The World-Wide Web. Communications of the ACM 1994;37(8):76.

4. Goldfarb C. SGML Handbook. New York: Oxford University Press, 1990.

5. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32(4):281.

6. Cimino JJ, Aguirre A, Johnson SB, Peng P. Generic queries for meeting clinical information needs. Bull Med Libr Assoc 1993;81(2):195.

7. Clyman JI, Powsner SM, Paton JA, Miller PL. Using a network menu and the UMLS Information Sources Map to facilitate access to online reference materials. Bull Med Libr Assoc 1993;81(2):207.

8. Cimino JJ, Johnson SB, Aguirre A, Roderer N, Clayton PD. The MEDLINE Button. Proc Annu Symp Comput Appl Med Care 1992:81.

9. Wiederhold G. Mediators in the architecture of future information systems. Computer 1992;25:38.

10. Fikes R, Engelmore R, Farquhar A, Pratt W. Network-based information brokers. Knowledge Systems Laboratory Report KSL-95-5, Stanford University, 1995.

11. Cimino C, Barnett GO. Analysis of physician questions in an ambulatory care setting. Comput Biomed Res 1992;25(4):366.

12. Powsner SM, Riely CA, Barwick KW, Morrow JS, Miller PL. Automated bibliographic retrieval based on current topics in hepatology: hepatopix. Comput Biomed Res 1989;22(6):552.

13. Purcell GP, Mar DD. SCOUT: information retrieval from full-text medical literature. Proc Annu Symp Comput Appl Med Care 1992:91.

14. Campbell JR, Kallenberg GA, Sherrick RC. The clinical utility of META: an analysis for hypertension. Proc Annu Symp Comput Appl Med Care 1992:397.

15. Horres MM, Starr SS, Renford BL. MELVYL MEDLINE: a library services perspective. Bull Med Libr Assoc 1991;79(3):309.

16. American National Standards Institute/NISO. Serial Issue and Contribution Identifier. Z39.56, 1991.